Machine Learning in Mineral Exploration: A Tutorial[1]

S. Ilkay Cevik (17sic1@queensu.ca)
Julian M. Ortiz (julian.ortiz@queensu.ca)

Abstract

This tutorial aims to demonstrate how to conduct some machine learning methods in geoscience to enhance mineral exploration studies. The document does not intent to present an exhaustive list of all the methods available currently, instead it aims to present some useful functions that exist in Python and R to conduct a workflow. The outputs of the procedures are presented hereby, and source codes are available in the notebooks of all the sections which are available on the GitHub page of the research group. They could be downloaded and repeated.

1. Introduction

This document is accompanied by a series of notebooks and aims to demonstrate how machine learning methods can enhance the geoscience studies, specifically mineral exploration activities. The data used in this study (raw data and final report) are publicly available (Geoscience BC, 2018). Geoscience BC, an independent, non – profit organization that generates earth science, re-analyzed almost 15,000 archived stream sediment samples with newer analysis methods from 20 1:250,000 NTS map sheet in northwest British Columbia. A private consultant company (CSA Global) was commissioned by Geoscience BC to compile and analyze the new dataset with classical and advanced statistical methods.

Stream sediment samples were used in this study therefore a sample does not represent a point, instead it represents all the geology inside of its watershed boundaries. For this reason, instead of working with the geological units, we work with geological terrains which are collections of geological units and covers wider areas. The description of the database is presented as a part of the exploratory data analysis tutorial. After exploratory data analysis, we present an application of supervised and unsupervised statistical methods to gain insight about the geology of the area. In the final part of the document, a discussion of the results is presented.

2. Exploratory Data Analysis (EDA)

Exploratory data analysis is the essential part of any statistical study. The EDA provides researchers with a chance to understand the structure of the data, detect and address the possible errors and decide what method is suitable to solve the problem that is being investigated.

In EDA, presenting summary statistics and plotting the data to see them are helpful. Therefore, we start by presenting summary statistics and plots. The following table is a summary table with some of the variables in the original data set. For brevity, only some part of it is presented here. Column names are

[1] Cite as: Cevik SI, Ortiz JM (2019) Machine Learning in Mineral Exploration: a Tutorial, Predictive Geometallurgy and Geostatistics Lab, Queen's University, Annual Report 2019, paper 2019-14, 10 p.

left as they are recorded in the database so that benefits of EDA can be seen clearly. Some field names are not self-explanatory, and EDA helps to understand their purpose.

Table 1 Summary statistics of the database. Note that this table does not contain all the features of the database. However, whole dataset can be found in the tutorial notebook.

	Count	**Unique**	**Top**	**Freq**	**Mean**	**Std**	**Min**	**0.25**	**0.5**	**0.75**	**Max**
MASTERID	15656	**15656**									
E_N83Z9_Val	15656				534499	134684	224897	437753	534591	636806	829089
N_N83Z9_Val	15656				6238924	167275	5874504	6111247	6224812	6387255	6548596
ORDR	15656	5	Primary	6583							
PHYS	15656	7	Youthful Mts	7352							
DRNP	15656	9	Dendritic	12459							
WTRC	15547	5	Colorless	11834							
SEDC	15628	14	Red	5809							
COMP	15644	38	2\2\0	3459							
DomGeo	15656	24	s_undivided	2705							
DomTerrain	15656	7	Stikinia	8126							
Au_ppb	15485				11.53	64.62	0.50	1.00	4.00	8.00	5400.00
LOI_pct	10869				7.60	6.39	0.05	3.60	6.00	9.70	77.50
Ag_ppm	**14874**				0.17	0.70	0.00	0.07	0.11	0.19	73.34
Pb_ppm	14874				9.68	16.38	0.29	4.87	7.07	10.10	1056.63
Zn_ppm	14874				95.59	76.59	6.00	60.50	84.10	110.10	1941.00

There are 15,656 samples in the database but only 14,874 of them have assay results of the elements. Those 782 samples without assay results are removed.

There are categorical variables as well as numerical. The following figures present the histograms of some of the elements and number of the samples that belong to each class in different categorical fields. It helps understanding the meaning of those fields. It is only clear after looking at Figure 1 that “DRNP” refers to drainage pattern, “PHYS” refers to physiography, “WTRC” represents the color of the water and “SEDC” represents the color of the sediment. It also shows that although there are 7 different geological terrains, only 4 of them comprise the total samples.

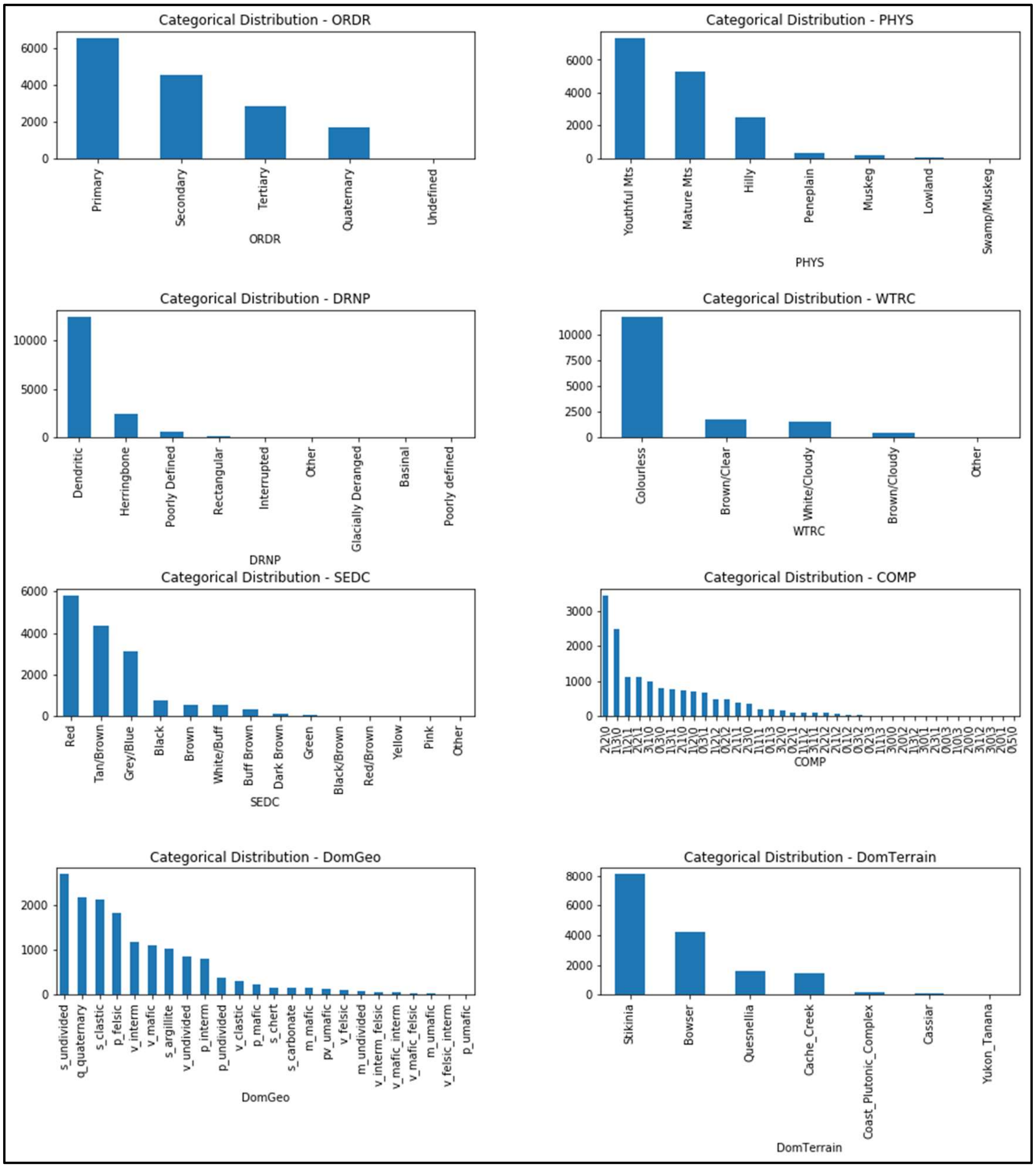

Figure 1 Distribution of the samples based on categories in different features.

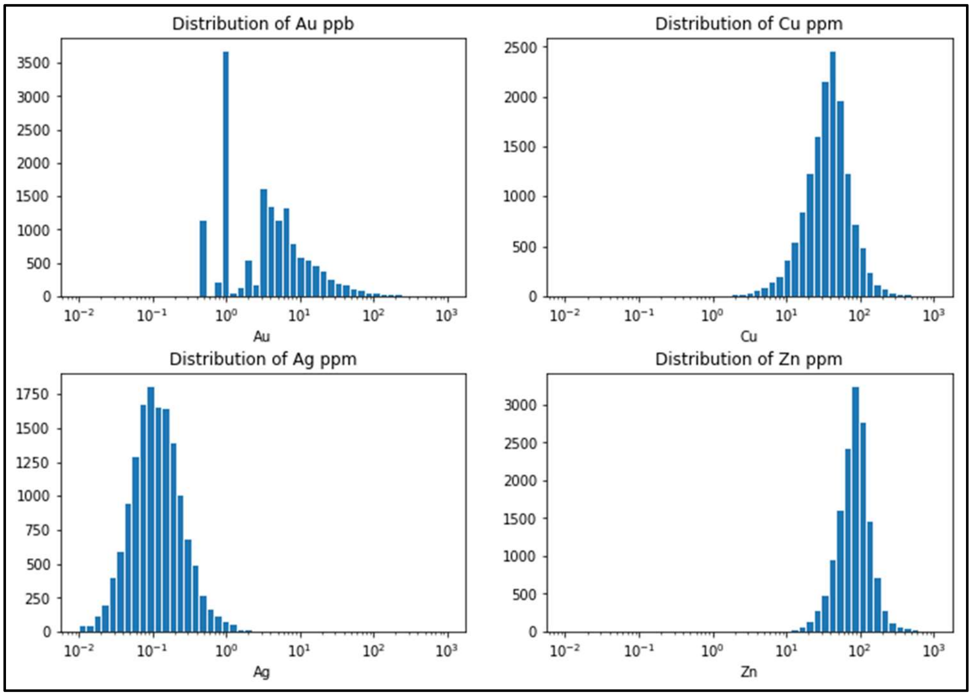

Figure 2 Distribution of the Au, Ag, Cu and Zn values

It is also worth checking the spatial distribution of the samples. Figure 3 shows the location of the samples with dominant terrain and assay results of copper.

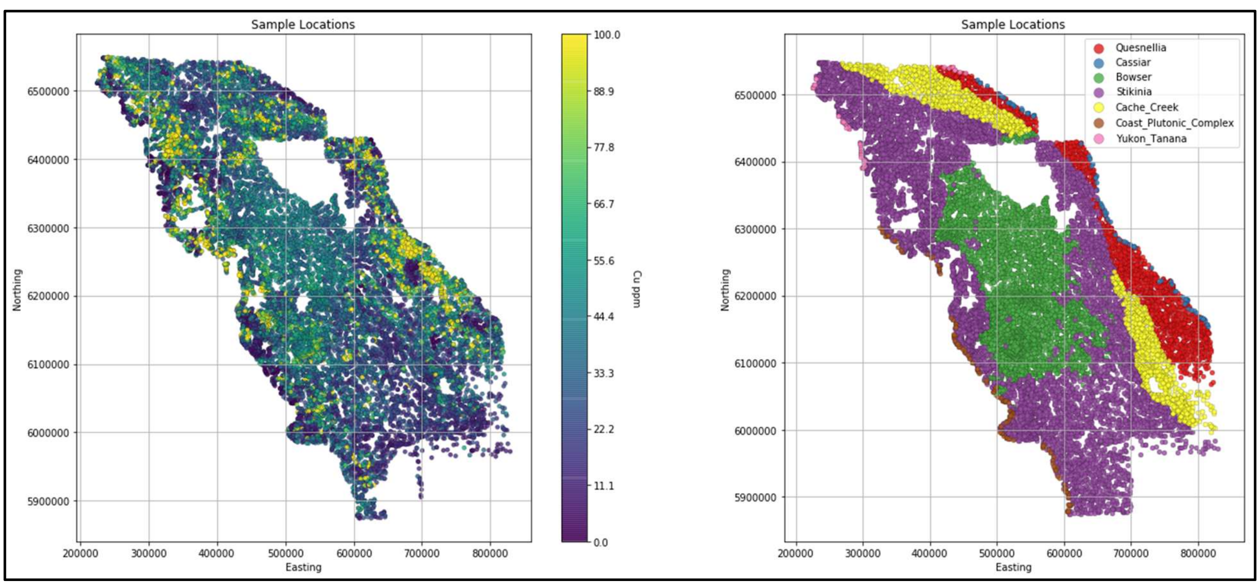

Figure 3 Location map of the samples with assay results of the copper (left) and dominant terrain colors (right)

From the figures, it seems that some of the units have consistently higher or lower copper values. In order to check the distribution of some elements in more detail, the following box and whisker plots are presented.

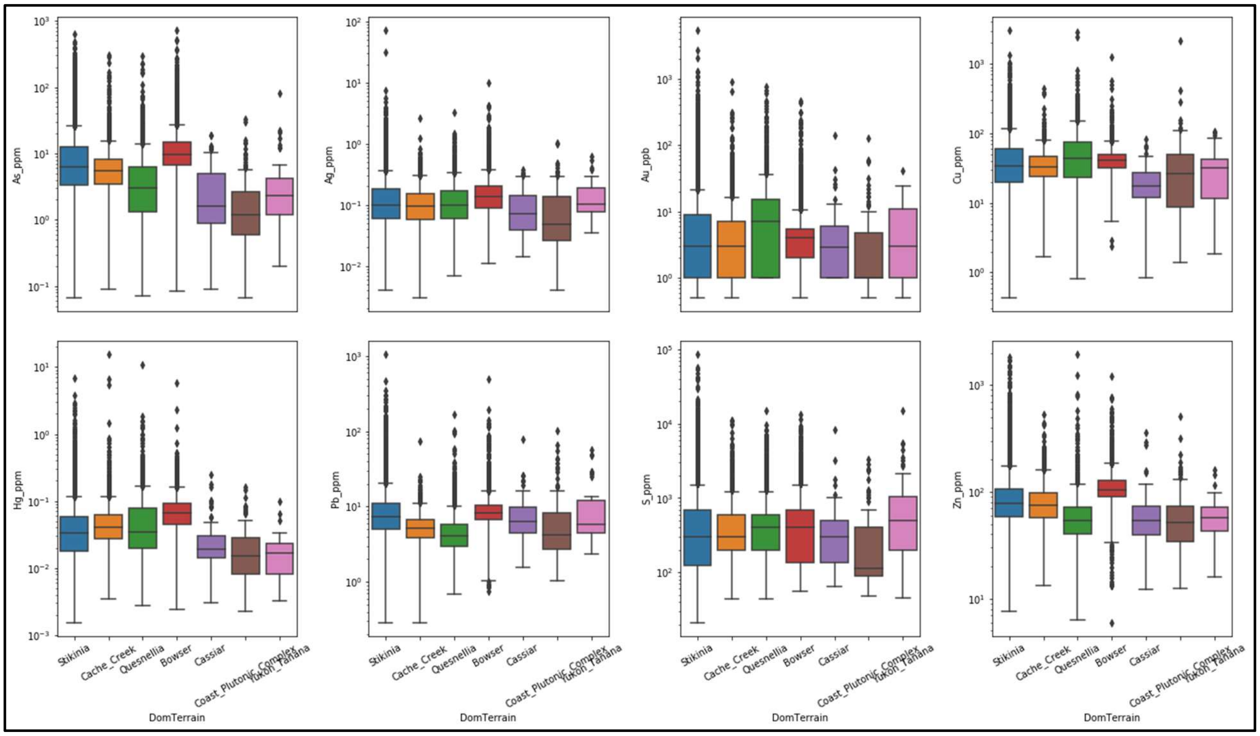

Figure 4 Box and whisker plots of some elements based on dominant terrain

EDA is an iterative process and it may be necessary to revisit some of the stages again after some more analysis. To summarize:

- Structure of the data and meaning of the features are understood
- Missing values are detected and removed from the database
- Spatial distribution of the sample points is presented

In the next section, a supervised classification method will be conducted.

3. Supervised Classification – Random Forest

In this section we will present the workflow to conduct a classification study by using geochemistry of the samples to predict their membership to a terrain. To present uncertainty in the prediction, 10 simulations are done for the prediction. A more robust study may include 100 simulations. This section starts with tuning the parameters of the algorithms.

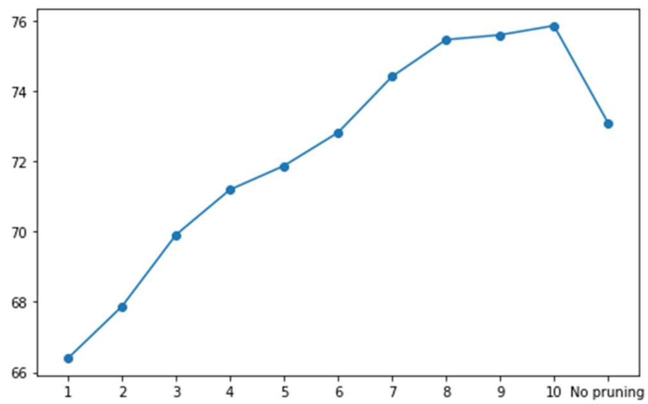

Figure 5 Prediction accuracy of the DT algorithm as a function of depth of tree.

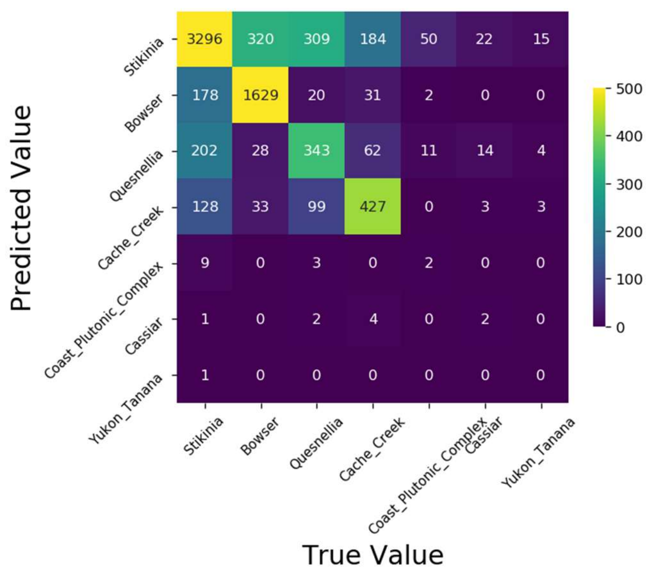

Figure 6 Confusion matrix of the DT prediction

Hyperparameter Tuning: In this section, hyperparameters of the decision tree (DT) and random forest (RF) are tuned. For DT, different depths for tree and for RF, different number of trees in the forest and number of variables to search for best split are tried, and results are presented below. 50% of the data set is kept for validation, however, internal error rate of Random Forest which is an internal unbiased error rate that random forest can calculate (out of bag (OOB) error rate), is also presented to compare.

By analyzing the sensitivity of DT, total depth of tree is chosen as 10 (Fig. 5) and 10 simulations of prediction are made. The distribution of the prediction accuracy is kept for future comparison with random forest results. In RF, there are 2 parameters which may have an impact on the prediction results. These are (1) the amount of the trees to grow in a forest and (2) the number of maximum features that is supposed to be selected randomly at each node to find the best split criterion among them. The following figures (Fig. 7) shows the prediction accuracy change of RF as a function of these two parameters.

The RF accuracy rate reaches a plateau with a total of 50 trees in the forest (Fig. 7). A total of 100 trees and the square root of the total variables to randomly choose in each split node are chosen for further analysis to get robust statistics about out of bag error rates.

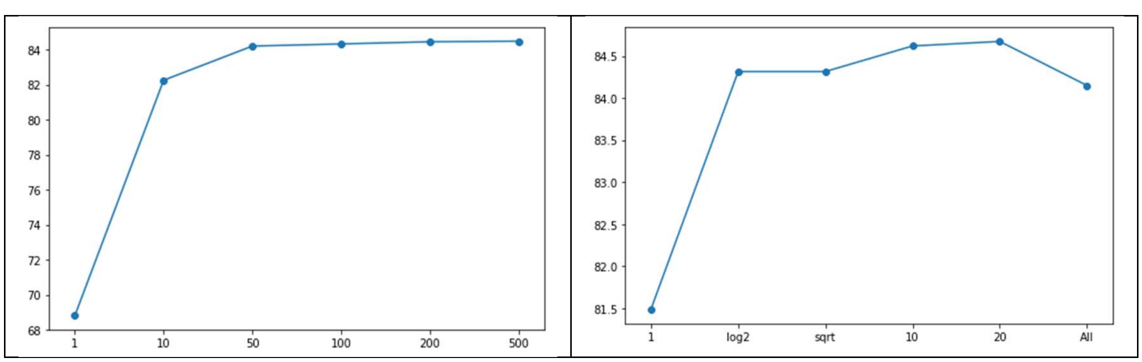

Figure 7 Prediction accuracy of RF as a function of number of trees in the forest (on the left) and number of variables to use in each split (on the right).

The average error rate of the validation set is **84.3%** while the internal out of bag error rate, which is calculated internally and does not require a validation set, of RF is **84%.** The confusion matrix of the last realization is presented below. Another useful feature of the RF is the fact that it calculates the feature importance internally. The result of the feature importance calculation after 10 simulations is presented as box plots to capture variability in the importance.

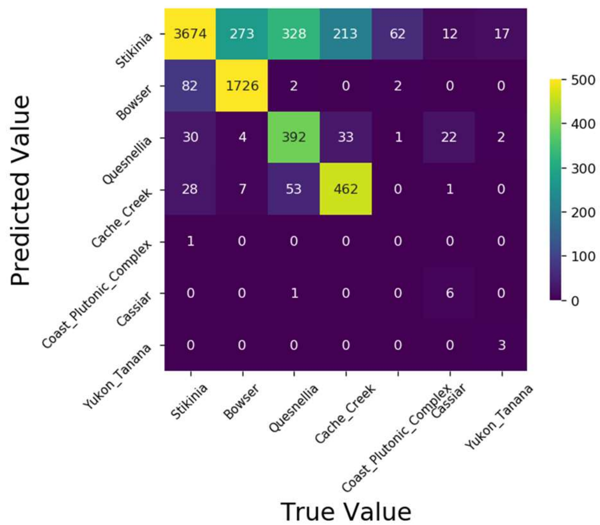

Figure 8 Confusion matrix of the RF classification

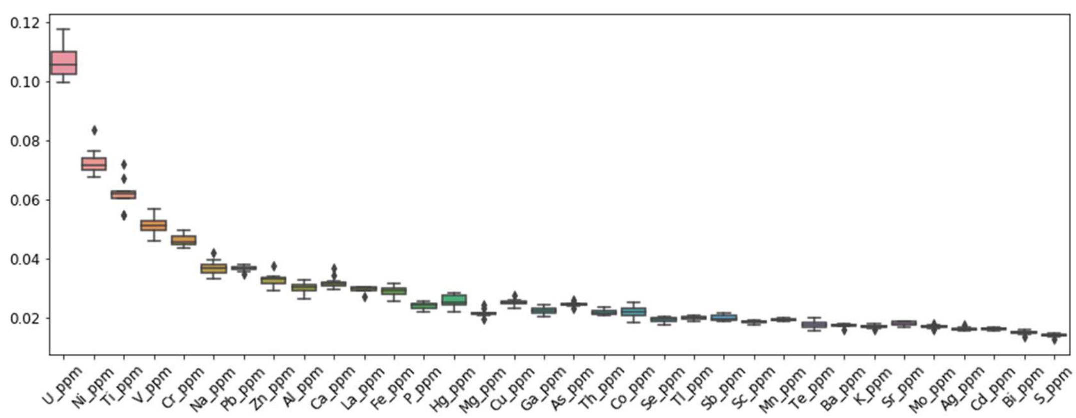

Figure 9 Feature importance of the RF classification after 10 realization.

To see the effect of using categorical variables, an R notebook is provided to do the same task in R since its random forest function is readily accepting categorical variables as inputs (Python can do this by creating new indicator variables for each class in a categorical feature). The OOB error rate of random forest reduced from **13.7%** to **11.4%** when, order of the stream (ORDR), physiography (PHYS), drainage pattern of the area (DRNP), color of the water (WTRC) and color or the sediments (SEDC) columns were added.

In a classification model, achieving accurate results does have many obvious predictive applications like producing geological maps, predictive models for mineralization, contamination However, having confusions also might indicate some useful information. In the maps below, we present an example for such a situation.

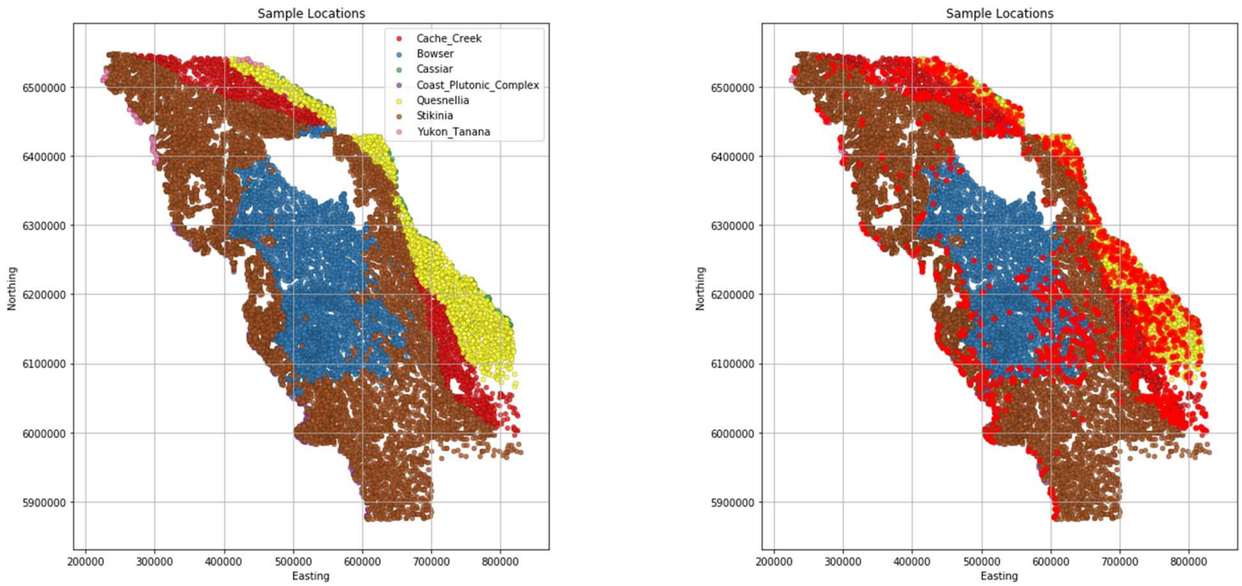

Figure 10 Location of the misclassified samples (Right hand side map – red samples)

Lithological information comes from interpretation of geologist and usually there is an uncertainty around the interpretation of the geologist, that is, one lithology can be interpreted in many ways by different geologists. Moreover, although training labels are used in deterministic ways, confusion in the algorithm may indicate that there might be an uncertainty of the phenomenon being investigated.

As it can be observed, misclassified samples are not distributed randomly, instead they are located close to the transition zones of different terrains (especially on the East side of the map). This supports the idea that misclassification may indicate places with an uncertainty in geology.

4. Unsupervised Clustering – Random Forest

Random Forest can be also used as a clustering tool with the help of its proximity matrix (Breiman 2001). This proximity matrix can be converted to a distance matrix which allows conducting clustering. This is an important feature since this kind of distance metric between samples is a result of assimilation of categorical and numerical data. The mechanism of the random forest clustering is introduced by Breiman 2001 and scrutinized by Shi and Horvath (2006). In this document, for the clustering section we use the R function produced by Shi and Horvath (2006).

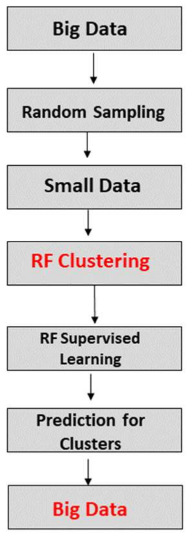

Figure 11
Clustering and classification workflow

One problem related to the random forest clustering is the fact that when the amount of the samples increases, the distance matrix generation process demands a lot of memory, as it creates an N x N matrix. In this tutorial, we present a method to overcome this problem by combining clustering with supervised learning. It simply consists of creating a smaller subset and conducting a clustering task in the small database. Once cluster labels are achieved, it becomes a classification problem. Therefore, cluster membership can be used as labels and a supervised classification can be conducted to assign all the samples to the clusters. Figure 11 summarizes this workflow. Related steps are presented as notebooks and can be repeated. The figures below show that RF clustering was successful to capture geologically sound clusters. The spatial distribution of the clusters coincides with either dominant geology and the dominant terrain maps although we conduct the test with very few samples (1000) and therefore, extrapolation by using RF supervised classifier was not highly accurate (80% accuracy rate).

More meaningful clusters can be achieved by increasing the number of samples and also increasing the number of trees and forests in the function. Additionally, dimension reduction methods can be used to reduce the amount of variables which will be used to generate the proximity matrix. Besides clustering directly the distance matrix, it is also promising to apply multidimensional scaling to the distance matrix to obtain coordinates of the samples. These coordinates, then, can be treated as additional variables to the original ones.

5. Conclusion

This tutorial presents that machine learning algorithms can be used to utilize a variety of data types in geoscience. The workflow presented is a way of assimilating categorical and numerical data which is an important feature for mineral exploration since geological field studies requires to be descriptive for the geologists. Both supervised and unsupervised methods presented here can be used for regional geochemical studies to delineate the main geological units or assess the uncertainty in a historic geological map to decide the focus of field studies. Their application can be extended to detect anomalies with the help of the same workflow.

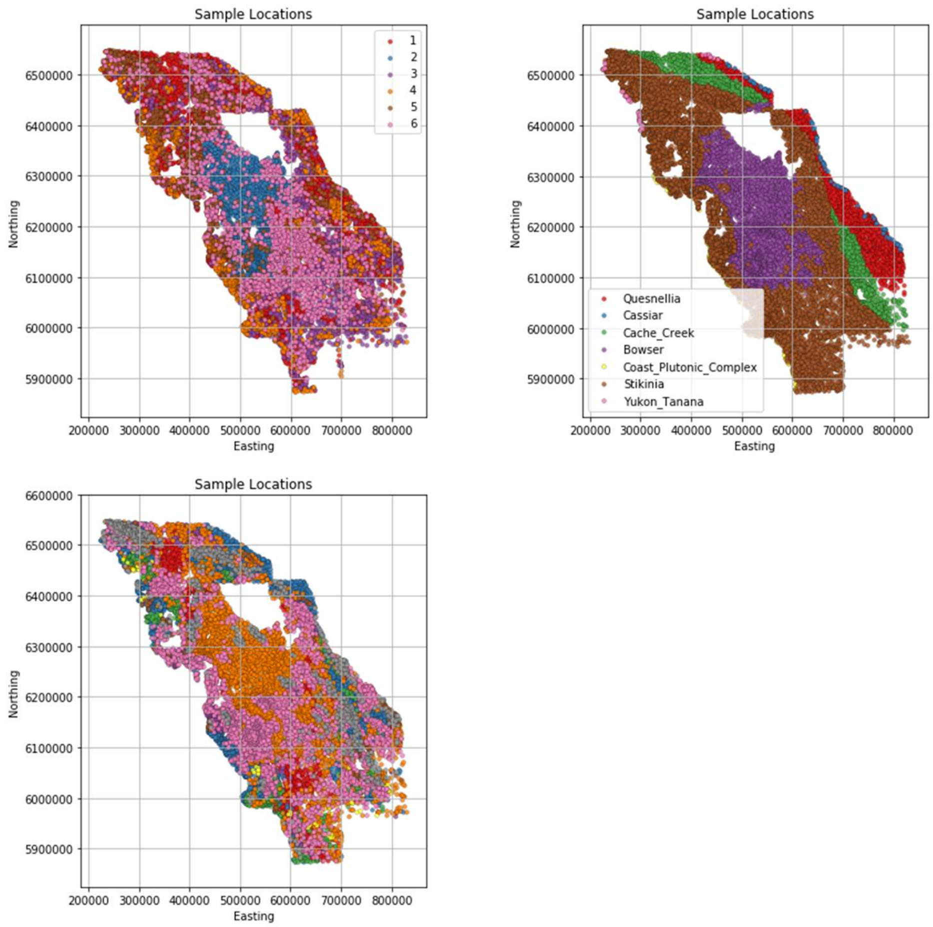

Figure 12 Location map of the RF clusters (Top left), dominant terrain (top right) and dominant geology (lower left)

6. References

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Geoscience BC, G. (2018, May 18). An Integrated Assessment of Regional Stream Sediment Geochemistry for Metallic Deposits in Northern Stikinia Terrane, British Columbia. Retrieved from http://www.geosciencebc.com/projects/2016-028/

Shi, T., & Horvath, S. (2006). Unsupervised learning with random forest predictors. Journal of Computational and Graphical Statistics, 15(1), 118-138.